A Survey Protocol to Assess Meaningfulness and Usefulness of Automated Topic Finding in the NASA Aviation Safety Reporting System

Carlos Paradis* and Rick Kazman†
University of Hawaii at Manoa, Honolulu, Hawaii, 96822, USA

Misty D. Davies‡ and Becky L. Hooey§
NASA Ames Research Center, Moffett Field, CA, 94035, USA

Context: The NASA Aviation Safety Reporting System (ASRS) is a voluntary confidential aviation safety reporting system. The ASRS receives reports from pilots, air traffic controllers, flight attendants and other involved in aviation operations. The reports are de-identified and coded by ASRS expert safety analysts and a short descriptive synopsis is written to describe the safety issue. The de-identified reports are then disseminated to the aviation community in a number of ways including entry into an online database, Safety Alert Bulletins and For Your Information Notices, and the CALLBACK newsletter. An opportunity of providing additional identification of safety concerns is with the use of topic modeling. Topic modeling can improve the dissemination of safety concerns by grouping and summarizing large collections of reports simultaneously. However, the generated summaries must be both meaningful and useful.

Aim: We propose a methodology to evaluate whether automated topic finding using topic modeling provides meaningful and useful topics.

Method: We extend the total error survey methodology to evaluate user topic comprehension of machine learning outputs. To accomplish this we performed a literature review to identify existing methods and define a construct for topic comprehension, utilizing existing ASRS synopsis writing practices to more precisely define meaningfulness and usefulness.

Results: Nine responses were obtained providing interpretations of computer-generated topics for evaluation. Participants interpretation of computer-generated topics of report sets, match the title and description of ASRS report sets written separately by analysts.

Conclusion: We conclude computer-generated topics, when grouping adjusted rand index is above 90%, are both meaningful and useful. However, the surveying of user understanding in machine learning outputs presents challenges due to the explosion of parameters to control for and the lack of systematic approach presented in the literature. More reproducible work and survey protocols are needed in the literature and our work is one step towards that direction.

I. Introduction

In NASA's Aviation Safety Reporting System (ASRS)*, each incident report is made publicly available, annotated with a one-line synopsis emphasizing the safety concerns and confounding factors. These synopses enable topic-driven exploration of individual records of safety concerns. In this work, we explore automating synopses for *groups* of records.

In prior work [1], we evaluated how well topic modeling can group reports by comparing automatically generated groupings against existing manually curated ASRS report sets. An ASRS report set† contains the 50 most recent reports at the time, grouped by safety topics that are commonly searched. The reports are selected and reviewed for relevance by ASRS analysts.

Topic modeling also provides a set of words which are intended to convey a synopsis of each grouping, commonly referred as "topics". Therefore, the topics obtained using topic modeling [2] are a suitable candidate for our goal. An

*Graduate Student, Department of Information & Computer Sciences.
†Professor, Shidler School of Business.
‡Research Computer Engineer, Intelligent Systems Division, Mail Stop 269-1, AIAA Associate Fellow.
§Director, NASA Aviation Safety Report System, P.O. Box 189.
*`https://asrs.arc.nasa.gov/`
†`https://asrs.arc.nasa.gov/search/reportsets.html`

example topic is shown in Table 1. In its "raw" form, topics are a list of words, where each word is assigned a probability. The table shows eight words, but a topic may contain thousands of words. The *display* is chosen a-posteriori by the analyst, to be presented to users. A common heuristic is to choose the 'top-n' terms with the highest probability for display using single terms.

Table 1 One topic and possible displays [3].

Topic Example		**Topic Display**
views	0.10	*Top Terms:* views, view, materialized, maintenance, warehouse, tables
view	0.10	
materialized	0.05	*Human:* materialized view, data warehouse
maintenance	0.05	
warehouse	0.03	*Single Term:* view, maintenance;
tables	0.02	*Phrase:* data warehouse, view maintenance
summary	0.02	*Sentence:* Materialized view selection and maintenance using multi-query optimization
updates	0.02	

Despite their potential applicability to generate a synopsis for a group of reports, the choice of the word "topic" in topic modeling exploits text-oriented intuitions, but no epistemological claims are actually made regarding these latent variables beyond their utility in representing probability distributions on sets of words [4]. Moreover, there is a longstanding assumption that the topics output by topic models are *meaningful* and *useful* [5, 6]. However, evaluating such assumptions is difficult because discovering topics is an unsupervised process. There is no gold-standard list of topics to compare against for every corpus, thus requiring exogenous data for evaluation [5].

We agree with [7] that topics themselves are not the end goal, but rather to use topics to improve some end-user task. Our end-user task is to provide meaningful and useful synopses for ASRS users. We therefore define the following research questions:

RQ1 Are assignments of topics to documents *meaningful*?

RQ2 Are assignments of topics to documents *usable*?

To answer our two research questions, we created a survey using the total survey error methodology [8], and defined, as gold standard, using existing ASRS report set data to evaluate if topics are meaningful and useful to serve as synopses for groups of aviation safety reports. For the existing synopses, we define *meaningfulness* and *usefulness* as follows:

Meaningfulness is defined as an aviation domain expert's ability to infer the meaning of an ASRS report set using only the computer-generated topics of the report set obtained from topic modeling.

Usefulness is related to the purpose of ASRS synopses. ASRS synopses are a short concise restatement of the primary safety concern and *if* clearly stated by the reporter, contributing factors might be noted. A synopsis typically starts with the reporter's job function "Air Traffic Controller reported that". To be *useful*, the topics provided by topic modeling summarizing groups of report should ideally include terms which suggest a) a safety threat and, b) contributing factors. The reporter's job function can be obtained separately from ASRS metadata, and is therefore not considered in our evaluation.

II. Method

Figure 1 provides an overview of the method, which is implemented in Kaona[‡]. We explain it briefly here as the method requires several steps, and we explore the details in the subsequent sections.

[‡]`http://github.com/sailuh/kaona`

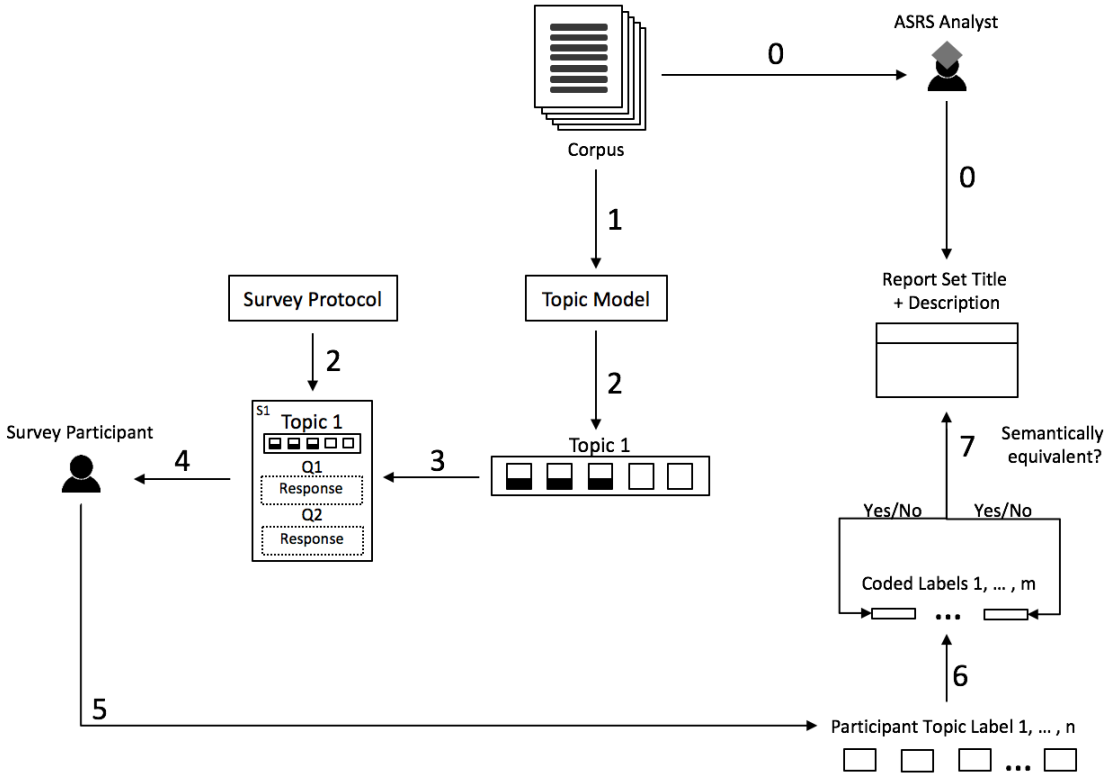


Fig. 1 Method Overview.

In our definition of *meaningfulness*, we stated that a domain expert must be able to infer a report set's meaning from a computer-generated topic. This means our method requires two paths: One to obtain a description of the report set to be used as a gold standard (step 0 in Figure 1), and a domain-expert interpretation of the topic (Figure 1 steps 1 through 5). If we have both, we can compare both to evaluate if they are semantically equivalent to answer our research questions.

The report sets title and description is already readily available on ASRS website. Hence our effort in this work is to obtain participant interpretations. We perform topic modeling using the report set's narratives (step 1, which does *not* include the title and description used as gold standard) to generate topics (step 2).

Next, to ask a participant their interpretation of a single topic, we used a survey protocol to define our questionnaire (step 2). In the questionnaire we present one topic at a time to a participant (step 4), to obtain one or more interpretations, i.e. one or more *participant topic labels* (step 5). We also asked participants to provide, for each topic label, the *computer generate topic's words* which influenced the topic label choice (step 5, not shown in the Figure). We group each participant's topic label + mapping into a *coded label* using open coding (step 6). Finally, we compared the *coded labels* to the rationale and description of step 0 to answer our research questions (step 7).

We now discuss in more detail each steps in the method overview, starting from the corpus.

A. Dataset

We used all 30 report sets from ASRS to generate topics for each report set, and chose two of the report sets for evaluation: a) Cabin Smoke, Fire, Fumes, or Odor Incidents, and b) General Aviation Flight Training Incidents. We chose these topics empirically, and we will explain how after we define our topic model setup in more detail. From each report in a report set we only used its *narrative*. That is, each report was treated as a unique document and used for topic modeling. We did not use the metadata or individual report synopses in this study. In addition to common stopwords [§], the following words were removed from the narratives as they would not convey meaning if occurring among the 'top-n terms': 'x', 'y', 'z', 'xx', 'yy', 'X', 'Y', 'Z', 'ZZZ', 'ZZZZ', 'zzz', 'zzzz', 'zzzzz', 'zzz1', 'zzz2', 'zzz3', 'zzz4', 'zz2',

[§]https://algs4.cs.princeton.edu/35applications/stopwords.txt

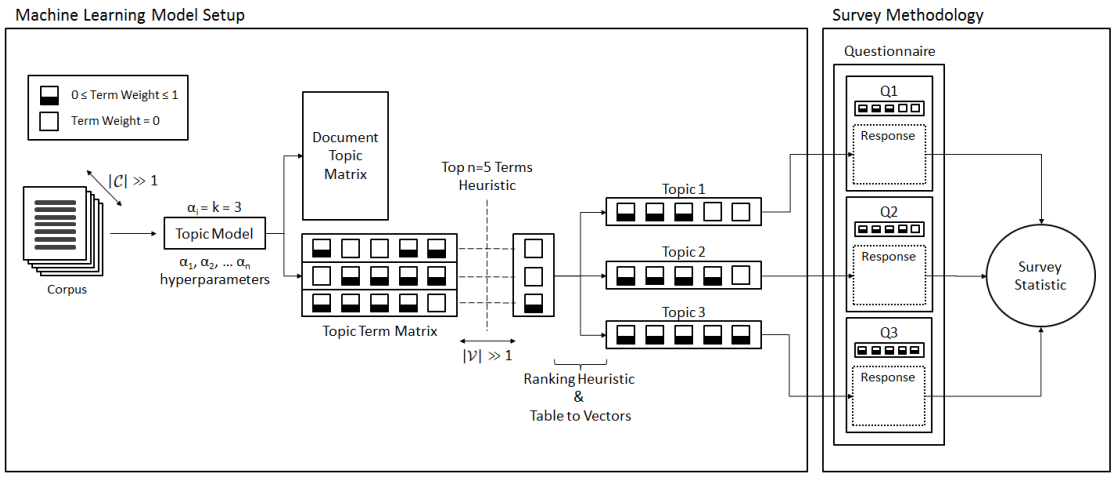


Fig. 2 Topic model setup and questionnaire steps. The choice of k = 3, n = 5 is for illustrative purposes only. We used k = 2 and n = 10 and maximum likelihood as ranking heuristic.

'zz3' (these are codes used by ASRS analysts to de-identify information such as airports, and navigation waypoints).

Each ASRS report set contains a title and a short description written by ASRS analysts that describes what the report set is about (note this is different from an individial ASRS report synopsis). We assume this information as our gold standard for the two chosen report sets. That is, to evaluate if the results are *meaningful*, our task is to compare if a survey response *is semantically equivalent* to the title and description of a report set (which is not curated by the authors of this paper), thus reducing subjectivity in answering our research question. The gold standard is not used in the corpus to train the algorithm.

B. Machine Learning Model Setup

Figure 2 provides further detail on steps 1, 2 and 3 of Figure 1. To evaluate if a topic generated from the corpus is meaningful and useful, we must ensure the topic obtained is from the same report set. Otherwise, the algorithm may include words in the topic from a different report set, despite the gold standard describing only a *single* report set at a time, which is an unfair comparison. Topic modeling algorithms, such as WarpLDA [9] used here, however, are primarily a grouping algorithm, which means that at least two groupings should be expected to exist in the data. To address this, we combined one pair of the report sets at a time to constitute the final corpus, and chose the pair which the topic model's output pair of groupings were the closest to the original pair of report sets (i.e. > 90% Adjusted Rand Index - ARI - [10]). For every possible pair, this results in two topics, one for each report set. In doing so, we controlled the evaluation of the topics to consider only the mechanism of choice of words in a topic, and not the performance of the grouping, the later which we defer to future work. The report sets topics with the highest ARIs were used in this study, which is how we empirically chose our 2 report sets indicated in the Dataset section.

Because we have 435 possible pair combinations of the 30 report sets (30C2), and we also chose k = 2 topics per report set, we have a total of 435*2 = 870 topics. This means that, for the same report set (e.g. fire and fumes), we will have 29 topics (sets of words), ranked in decreasing ARI order, indicating how well the topic model was able to recover the original report sets.

To exemplify the rationale, consider two pairs of the 435 possible pair combinations that the titles are: a) Cabin Smoke, Fire, Fumes, or Odor Incidents **X** Controller Reports, and b) Cabin Smoke, Fire, Fumes, or Odor Incidents **X** Passenger Electronic Devices. The topics obtained for Cabin Smoke, Fire, Fumes, or Odor Incidents in a) and b) are as follows:

- a) smoke | cabin | fire | odor | captain | passengers |fumes | cockpit | maintenance | attendant
  - (100% ARI on report set and topic modeling grouping match)
- b) captain | odor | cockpit | smelled | atc | crew | fumes | maintenance | fa | officer
  - (30% ARI report set and topic modeling grouping match)

As we can see and would expect, the topic model groupings that more closely approximates the original pair of report sets more closely convey its original title. Our interest in this work is to assess if the better groupings are meaningful

and useful. Our work setup is the same as [1], except we do not perform stemming here. This is because the unstemmed terms generally result in topics that are more easily interpreted [7], which is what we seek to assess in this work.

Lastly, since each participant was expected to spend no more than one hour in the questionnaire, a total of 5 set of words were presented to each participant; here we present 2 of them. Using the criteria above, the report sets, and the set of words derived from them were:

1) **Cabin Smoke, Fire, Fumes, or Odor Incidents.** smoke | cabin | fire | odor | captain | passengers | fumes | cockpit | maintenance | attendant
2) **General Aviation Flight Training Incidents.** tower | pilot | feet | traffic | plane | student | pattern | turn | airport | downwind

These 2 topics are represented in Step 2 of Figure 1. Further details on the setup may be found in our prior work [1], in the interest of focusing on the survey method which is the novel portion of this paper.

C. Survey Method

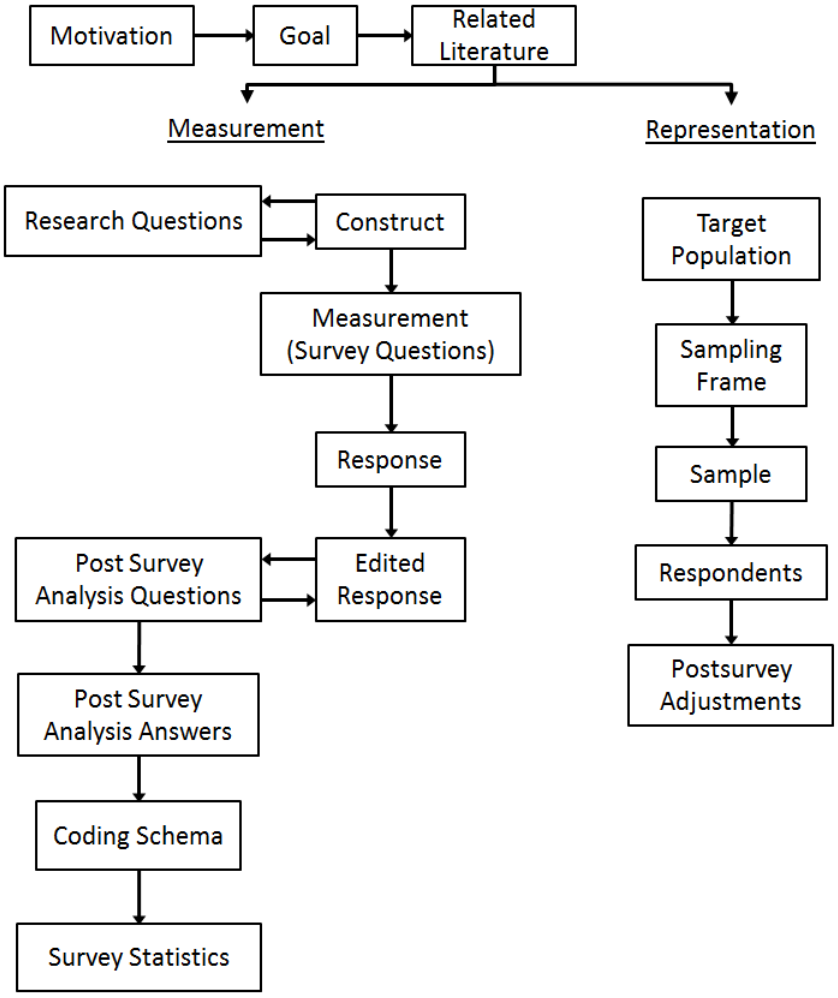


Fig. 3 Adapted from the Total Error Survey Design Method [8].

In this section, we present how we surveyed subject matter experts to assess if the topics were meaningful and useful. To do so, we defined a survey protocol, as shown on step 2 of Figure 1. Based on [8], a survey protocol must establish a set of steps, to be reproducible and minimize error. To account for the machine learning setup in our work, we adapted the survey protocol method steps as shown in Figure 3). The method branches in two independent steps to design the survey (left) , and define the representation (right). Each step of the figure is defined in the following subsections.

1. Motivation, Goal and Related Literature

Our motivation is to enable faster exploration of common safety themes. Ideally, comprehension would only require a set of words (as opposed to forcing an analyst to read a large number of reports). This would enable readers to select or skip entire groups of documents when searching for safety threats. Our goal is therefore to assess if the sets of words are both *meaningful* and *useful* to readers. To realize our goal we conducted a literature review to inform our survey decisions, in particular the construct shown in Figure 4.

2. Research Questions and Construct

We described our construct, the elements of information that are sought by us [8, p.41] in Figure 4. We can see the topic construct is broken down into 3 parts: *usefulness*, *display* and *topic comprehension* (meaningfulness). In the following subsections, we will explore each of the 3 components.

The goal of choosing a *display* for topics in a topic-term matrix is to improve *topic comprehension*. Specifically, a good display will allow a reader to identify a *topic label* from a pool of terms or from the person's own vocabulary, be it a single-term hypernym or phrase. There are various forms of *display*, but the literature does not discuss their differences, which makes it hard to assess topic comprehension as a whole. Our topic construct serves as a way to reason about the related literature from a topic comprehension standpoint. It is necessary, but not sufficient that a topic is comprehended through the terms. The comprehended meaning must correctly represent the underlying documents. That is, if the topic label is seen as a concept, then the documents it is assigned are seen as the concept's expression.

3. Measurement (Survey Questions)

For our survey measures, we did not use graphics, only words. We defer to future work the assessment of displays using *visual components* [12], such as representing topics similarity in a 2D plane, and/or the use of *metadata*. Our measurement focuses on *ranked terms*, and the associated *ranked heuristics* (as commonly reported in the literature, but not empirically validated).

Before we describe the questions we presented to survey participants, we first present some context on the origin of the sets of words. In providing context, it was important that our survey not introduce biases in the responses when we attempted to explain what we mean by topic *meaningfulness*. To do so, we borrowed the explanation from [7], with some modifications. The following shows our changes compared to the original (strike through represents removed text, bold represents added text, and normal text represents original text in the cited work):

> The ~~topics learned~~ **set of words displayed** by a topic model **(a computer program which automatically groups similar documents based on their content and provide set of words to describe them)**, are usually sensible, meaningful, interpretable and coherent. But some ~~topics learned~~**displayed sets of words** (while statistically reasonable) are not particularly ~~useful~~**meaningful** for human use. To evaluate our methods, we would like your judgment on how "~~useful~~**meaningful**" some ~~learned topics~~ **sets of terms** are. Here, we are purposefully vague about what is "~~useful~~**meaningful**" ... it is some combination of coherent~~, meaningful~~, interpretable, words-are-related, subject-heading-like, something-you-could-easily-label, etc.

First, we replaced the term 'useful' from the original question with 'meaningful'. We made this change, as we believe it more accurately captures the intent of the question and because we have a more precise definition of 'useful' leveraging the domain. Second, we explained the *display* in simpler terminology, as it is not necessarily the case that our target population is familiar with the method, nor is such knowledge relevant to our goal.

Provided with this context, we asked participants to write, if possible, in their response to the first question one or more topic labels out of the set of words provided (right side of Figure 2). In the second question, we ask participants to provide a mapping from the topic labels to the original set of words, so we are able to assess consensus between topic label choices, even if the choice of words are not the exact same. Because our questions and response formats are closely related, we provide the exact question after describing the response type next.

4. Responses

As shown earlier in Table 1, the choice of topic label display *using text* can vary (e.g. single term, phrase, sentence). In the Table, the 'Human:' topic labels included an inferred word, *data*, i.e. the word was not provided by the computer, but guessed by a person using domain knowledge. And the topic label itself could be an entire sentence. The topic labels presented in the Table may be derived from a subset of the set of words provided, in which case some of the

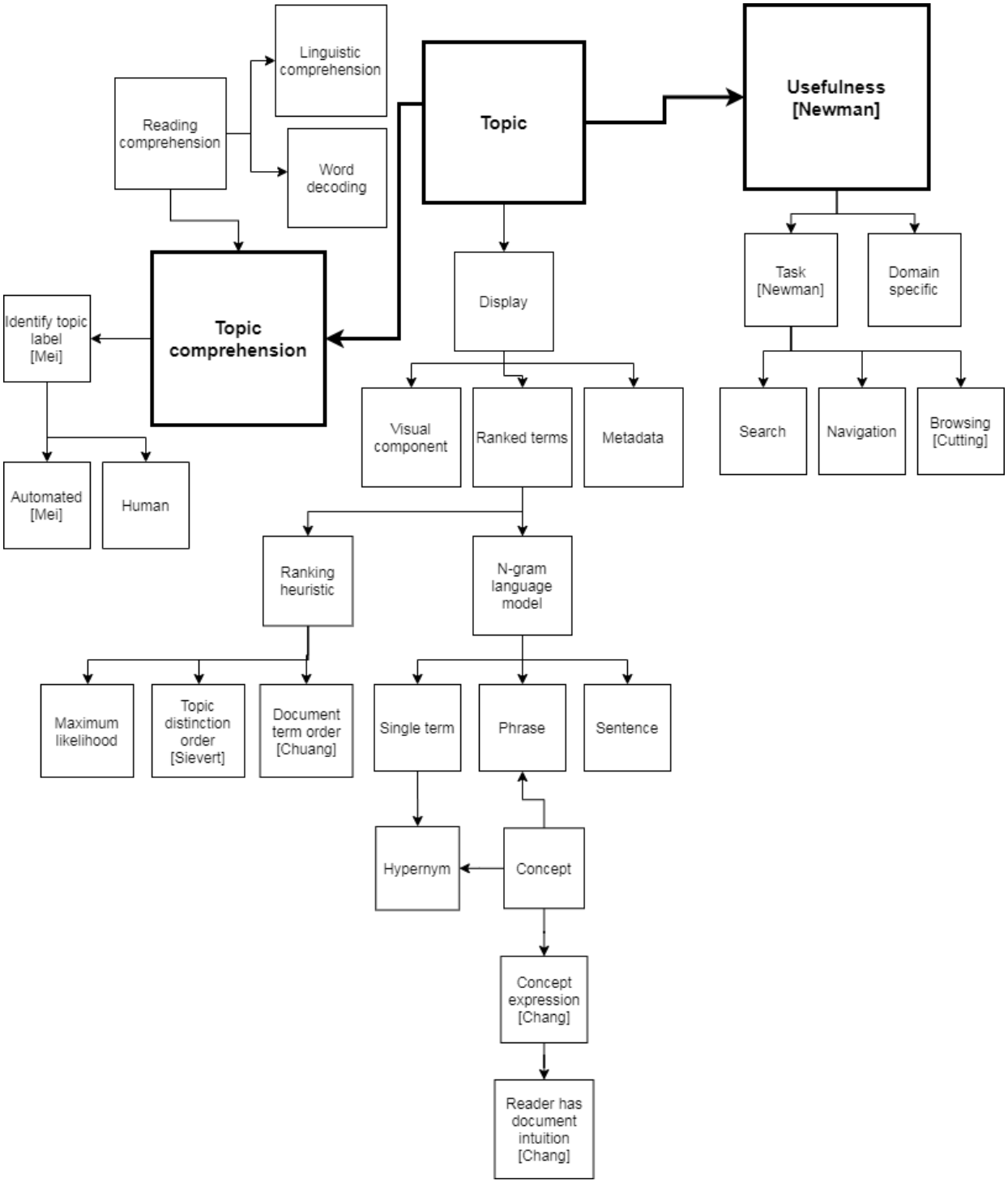


Fig. 4 Topic Construct [3–7, 11–20].

words in the topic label end up being "noise". Ideally, per the *ranking heuristic* of maximum likelihood, these words would not occur among the top n terms.

What we see from this example is that the choice of topic label is a process rich with decisions and implicit assumptions, which could be missed in an overly structured response, such as multiple choice. We also hypothesize that the collection of those choices also relates to a participant's ability to derive meaning from the set of words presented, and also challenges existing assumptions in the related literature. For example, [3] assumes that the use of *phrases* leads to more meaningful topic labels, which in turn is used as a replacement to present topics to users instead of a set of words. When we chose to present a set of words instead of, for example, a set of randomly selected report snippets [14], we also made an assumption about topic display, similar to one of the seminal works in topic modeling [4]. Since no single response type seem to suffice to capture these decisions, we decided to make our two questions open-ended.

While the survey response was open-ended, the response was not fully unstructured. We felt that not providing any guidance (beyond what type of answer was expected) would severely limit the comparison of responses and subsequent analysis. To address this limitation, we defined the response similar to [3] and exemplified in Table 1. The full survey is provided in the appendix.

For clarity, and similar to [7], we also included an example response. However, to be conservative we decided to use template example answers, as shown above (i.e. word1 word5) instead of actual words to avoid introducing bias. Finally, we asked participants to note the time of the start and end of their responses, to assess the difficulty in interpreting the set of words as a proxy measure of topic comprehension.

5. Post Survey Analysis Question and Answers

Thus far, we have presented *how* we instructed participants to answer our research questions, but we did not explain *why* these instructions were adequate. We included two steps to the total survey error measurement methodology [8], the *post survey analysis question*, and the *post survey analysis answer*, to make our rationale more explicit.

If the Measurement (Questions Section) and response instructions (Response Section) serve as a process to generate data from the participants, i.e. the actual responses, then the post-survey analysis question and post survey analysis answer sections explicitly define what we intend to analyze from the responses as a set of questions to the responses (post survey analysis questions), and the answer to these inquires (post survey analysis answers). Indeed, the measurement and response sections were guided by the post survey analysis questions, which in turn were based on related literature. The post survey analysis answers were encoded using the Coding Schema, as defined in the next section, and then evaluated with survey statistics to answer our research questions.

Our post survey analysis questions, answers and the associated rationale of why they help us answer the research questions are as follows:

- RQ1. Is the assignment of topics to documents meaningful?
  - **Post Survey Analysis Question 1.** Are the **coded labels from participant's topic labels** semantically related to the report set title and description from which the set of words were derived?
    - **Post Survey Analysis Answer.** A vector of Yes/No responses, where the number of elements of the vector equals the number of topic labels.
    - **Rationale.** While a participant may infer any number of topic labels (meaningfulness), the inferred topic labels still have to correctly represent the set of underlying documents of the topic, which is reflected on the report set title and description the topics were derived from.
- RQ2. Is the assignment of topics to documents useful?
  - **Post Survey Analysis Question 2.** Are **coded labels from participant's topic labels** semantically related to the report set title and description from which the set of words were derived *and* and related to an aviation safety issue ?
    - **Post Survey Analysis Answer.** A vector of Yes/No responses, where the number of elements of the vector equals the number of topic labels.
    - **Rationale.** While a participant may infer any number of topic labels (meaningfulness), which are safety threats (usefulness), the inferred topic label still has to correctly represent the set of underlying documents of the topic (i.e. a participant inferring an incorrect safety issue would not be a meaningful assignment).

6. Coding Schema

Since the post survey analysis answers are Yes/No vectors, a vector of 1/0s quantifies meaningfulness and usefulness for our post survey analysis questions.

7. Target Population

Our target population are ASRS Analysts and Pilots, as both are familiar with the ASRS database and reports.

8. Sampling Frame

The frame population is the set of target population members that has a chance to be selected into the survey sample (e.g. the sampling frame of a target population of U.S. adults could be a list of telephone numbers) [8, p.45]. In this study, they were either NASA employees, or affiliates on NASA contracts. We refer to them as subject-matter experts (SMEs). Many of them have experience across several domains of aviation, including air traffic control, dispatch, etc.

9. Sample

We use convenience sampling on the sampling frame. We have asked contacts who have helped us in the past to volunteer as respondents.

10. Respondents

A total of 13 people volunteered to participate in the survey, two of which did not follow-up. We first presented two participants with the survey for prototyping, and and the remainder nine to take the final version of the survey.

11. Postsurvey Adjustments

We manually transcribed all responses to the two surveys questions to a tabular format, modifying only grammar errors on words. The tables created contained the following columns:

- Raw Table 1
  - Participant ID
  - Survey ID
  - Participant Topic Label
  - Term

Participants could have multiple topic labels and multiple terms per topic label. Topic labels which did not include set of words associated to them were eliminated from the table.

12. Survey Statistics

We compared the provided topic labels to the gold standard (title and description of the report set provided by ASRS) to assess if topics were meaningful and useful.

III. Results

Figures 5 and 6 show participant topic labels, the coded labels, and the report set title and description. The mapping of each topic label to a topic's word provided by a participant was colored and organized to show related themes. We also provide our mapping from the participant's topic labels to our coded labels based on the information provided.

RQ1. Is the assignment of topics to documents meaningful?

Post Survey Analysis Question 1. Are the **coded labels from participants' topic labels** semantically related to the report set title and description from which the set of words were derived?

In this and the following RQ, we evaluate if the identified coded labels reflect the underlying reports from which the set of words were derived. Since examining every report in ASRS would be more subjective, we compared the coded labels instead to each report set's title and abstract (shown in Figures 5 and 6 to the left). Note in this RQ our interest is to see if the coded labels are related to the title or abstract, *regardless* of being safety threat related.

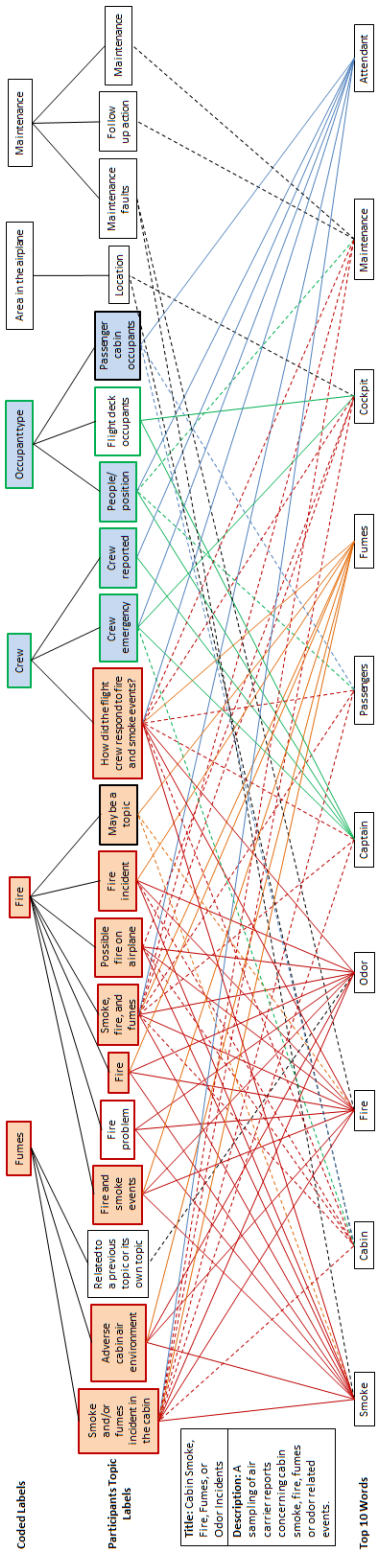


Fig. 5 Report Set: Cabin Smoke, Fire, Fumes, or Odor Incidents.

We can see across all report sets in Figures Figure 5 and 6, some of the coded labels relate exactly to the main theme of the report set, which is highlighted by the colors in the coded labels. In Figure 5, the coded labels “fire” and “fumes”

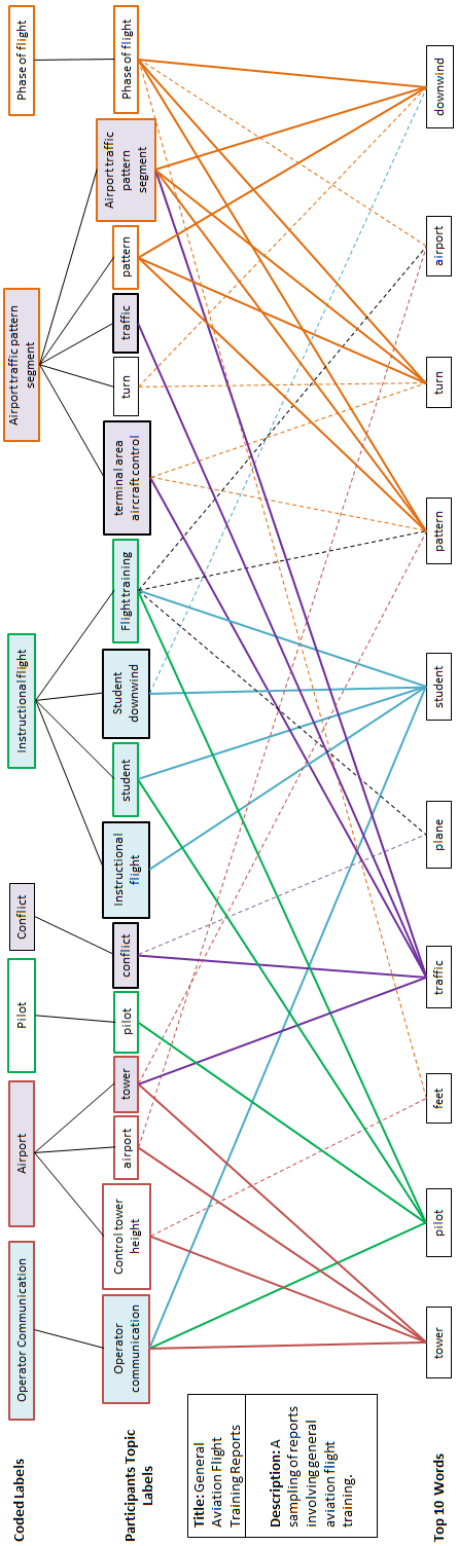


Fig. 6 Report Set: General Aviation Flight Training Incidents.

relate to the title and description of the report set, Cabin Smoke, Fire, Fumes or Odor Incidents. This is expected, given the set of words generated by the algorithm also contain these words and others (e.g. odor). The coded labels

in Figure 6, although not related to safety threats, properly relate to the report set meaning. In Figure 6, the coded label "instructional flight", is associated to the report set of training reports. We can see in particular the report set of instructional flight in Figure 6 that a single word led to the proper inference of the report set, i.e. 'student'.

The answer to RQ1 is therefore *yes*, participants were able to interpret, from the set of words, the meaning of the report sets.

RQ2. Is the assignment of topics to documents useful?

Post Survey Analysis Question 2. Are **coded labels from participants' topic labels** semantically related to the report set title and description from which the set of words were derived *and* to safety issues?

From Figure 5, we can see the report set conveys a safety issue. Figure 6 does not convey a safety threat, but the report set itself also does not. We therefore conclude that the answer to RQ2 is yes.

IV. Conclusion and Future Work

In this work, we proposed a survey protocol to assess if topics provided by topic modeling are meaningful and useful in the NASA Aviation Safety Reporting System. Unlike prior work, we presented a survey protocol that can be reused in other topic modeling experiments. Contrary to standard surveys, we found protocols that evaluate machine learning outputs require additional consideration on their presentation to survey participants so that conclusions can be generalized. Based on the results, we concluded that topics can be used to summarize report sets which contain specialized vocabulary. In future work, we plan to assess how the performance of grouping affects the topics meaningfulness and usefulness.

Acknowledgments

The material is based upon work supported by NASA under award No 80NSSC19M0202. This research was partially conducted at NASA Ames Research Center. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement by the United States Government.

References

[1] Paradis, C., Kazman, R., Davies, M., and Hooey, B., "Augmenting Topic Finding in the NASA Aviation Safety Reporting System using Topic Modeling," 2021. https://doi.org/10.2514/6.2021-1981, URL https://arc.aiaa.org/doi/abs/10.2514/6.2021-1981.

[2] Binkley, D., Heinz, D., Lawrie, D., and Overfelt, J., "Understanding LDA in Source Code Analysis," Proceedings of the 22nd International Conference on Program Comprehension, Association for Computing Machinery, New York, NY, USA, 2014, p. 26–36. https://doi.org/10.1145/2597008.2597150, URL https://doi.org/10.1145/2597008.2597150.

[3] Mei, Q., Shen, X., and Zhai, C., "Automatic Labeling of Multinomial Topic Models," Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, USA, 2007, p. 490–499. https://doi.org/10.1145/1281192.1281246, URL https://doi.org/10.1145/1281192.1281246.

[4] Blei, D. M., Ng, A. Y., Jordan, M. I., and Lafferty, J., "Latent dirichlet allocation," Journal of Machine Learning Research, Vol. 3, 2003.

[5] Chang, J., Gerrish, S., Wang, C., Boyd-graber, J. L., and Blei, D. M., "Reading Tea Leaves: How Humans Interpret Topic Models," Advances in Neural Information Processing Systems 22, edited by Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, Curran Associates, Inc., 2009, pp. 288–296. URL http://papers.nips.cc/paper/3700-reading-tea-leaves-how-humans-interpret-topic-models.pdf.

[6] Blei, D. M., "Probabilistic Topic Models," Commun. ACM, Vol. 55, No. 4, 2012, p. 77–84. https://doi.org/10.1145/2133806.2133826, URL https://doi.org/10.1145/2133806.2133826.

[7] Newman, D., Noh, Y., Talley, E., Karimi, S., and Baldwin, T., "Evaluating Topic Models for Digital Libraries," Proceedings of the 10th Annual Joint Conference on Digital Libraries, Association for Computing Machinery, New York, NY, USA, 2010, p. 215–224. https://doi.org/10.1145/1816123.1816156, URL https://doi.org/10.1145/1816123.1816156.

[8] Groves, R. M., Jr., F. J. F., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R., Survey Methodology, 2nd ed., Wiley, 2009.

[9] Chen, J., Li, K., Zhu, J., and Chen, W., "WarpLDA: a Simple and Efficient O(1) Algorithm for Latent Dirichlet Allocation." CoRR, Vol. abs/1510.08628, 2015. URL http://dblp.uni-trier.de/db/journals/corr/corr1510.html#0001LZC15.

[10] Vinh, N. X., Epps, J., and Bailey, J., "Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance," J. Mach. Learn. Res., Vol. 11, 2010, p. 2837–2854.

[11] Chuang, J., Manning, C. D., and Heer, J., "Termite: Visualization Techniques for Assessing Textual Topic Models," Proceedings of the International Working Conference on Advanced Visual Interfaces, Association for Computing Machinery, New York, NY, USA, 2012, p. 74–77. https://doi.org/10.1145/2254556.2254572, URL https://doi.org/10.1145/2254556.2254572.

[12] Smith, A., Malik, S., and Shneiderman, B., Visual Analysis of Topical Evolution in Unstructured Text: Design and Evaluation of TopicFlow, Springer, 2015, pp. 159–175. https://doi.org/10.1007/978-3-319-19003-7_9, URL https://doi.org/10.1007/978-3-319-19003-7_9.

[13] Sievert, C., and Shirley, K., "LDAvis: A method for visualizing and interpreting topics," 2014. https://doi.org/10.13140/2.1.1394.3043.

[14] Cutting, D. R., Karger, D. R., Pedersen, J. O., and Tukey, J. W., "Scatter/Gather: A Cluster-Based Approach to Browsing Large Document Collections," Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Association for Computing Machinery, New York, NY, USA, 1992, p. 318–329. https://doi.org/10.1145/133160.133214, URL https://doi.org/10.1145/133160.133214.

[15] Newman, D., Lau, J. H., Grieser, K., and Baldwin, T., "Automatic Evaluation of Topic Coherence," Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, USA, 2010, p. 100–108.

[16] Mimno, D., Wallach, H. M., Talley, E., Leenders, M., and McCallum, A., "Optimizing Semantic Coherence in Topic Models," Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, USA, 2011, p. 262–272.

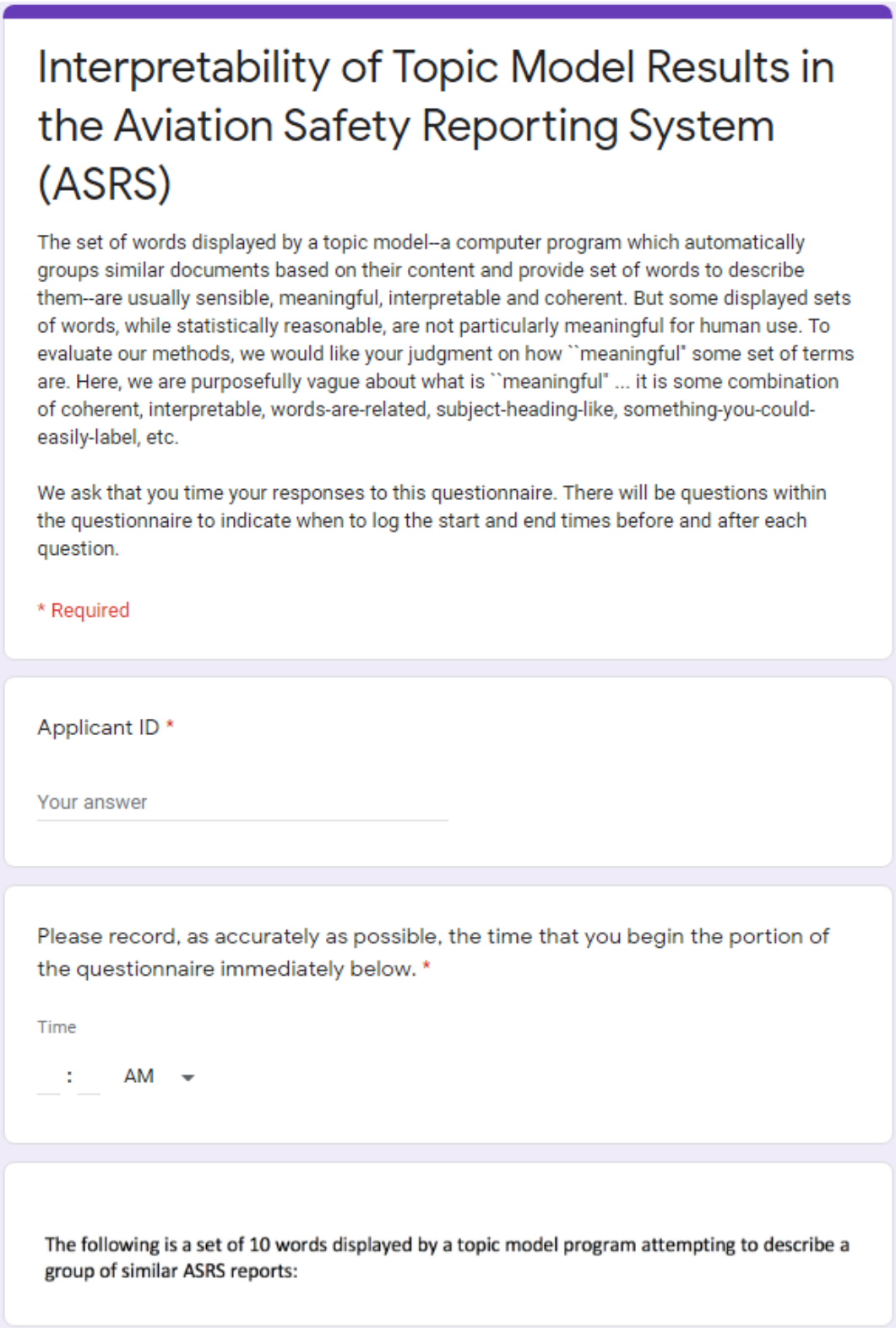
[17] AlSumait, L., Barbará, D., Gentle, J., and Domeniconi, C., "Topic Significance Ranking of LDA Generative Models," Machine Learning and Knowledge Discovery in Databases, edited by W. Buntine, M. Grobelnik, D. Mladenić, and J. Shawe-Taylor, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009, pp. 67–82.

[18] Gough, P. B., and Tunmer, W. E., "Decoding, Reading, and Reading Disability," Remedial and Special Education, Vol. 7, No. 1, 1986, pp. 6–10. https://doi.org/10.1177/074193258600700104, URL https://doi.org/10.1177/074193258600700104.

[19] Oakhill, J., Cain, K., and Bryant, P., "The dissociation of word reading and text comprehension: Evidence from component skills," Language and Cognitive Processes, Vol. 18, No. 4, 2003, pp. 443–468. https://doi.org/10.1080/01690960344000008, URL https://doi.org/10.1080/01690960344000008.

[20] Röder, M., Both, A., and Hinneburg, A., "Exploring the Space of Topic Coherence Measures," Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, Association for Computing Machinery, New York, NY, USA, 2015, p. 399–408. https://doi.org/10.1145/2684822.2685324, URL https://doi.org/10.1145/2684822.2685324.

A. Questionnaire



Interpretability of Topic Model Results in the Aviation Safety Reporting System (ASRS)

The set of words displayed by a topic model--a computer program which automatically groups similar documents based on their content and provide set of words to describe them--are usually sensible, meaningful, interpretable and coherent. But some displayed sets of words, while statistically reasonable, are not particularly meaningful for human use. To evaluate our methods, we would like your judgment on how ``meaningful" some set of terms are. Here, we are purposefully vague about what is ``meaningful" ... it is some combination of coherent, interpretable, words-are-related, subject-heading-like, something-you-could-easily-label, etc.

We ask that you time your responses to this questionnaire. There will be questions within the questionnaire to indicate when to log the start and end times before and after each question.

* Required

Applicant ID *

Your answer

Please record, as accurately as possible, the time that you begin the portion of the questionnaire immediately below. *

Time

: AM

The following is a set of 10 words displayed by a topic model program attempting to describe a group of similar ASRS reports:

Fig. 7 Survey v4 p1.

The following is a set of 10 words displayed by a topic model program attempting to describe a group of similar ASRS reports:

smoke | cabin | fire | odor | captain | passengers | fumes | cockpit | maintenance | attendant

format: word1 | word2 | word3 | word4 | word5 | word6 | word7 | word8 | word9 | word10

1. Please **identify**, if possible, in the following answer one or more **topic labels** that you believe the set of words are about. For each topic label, you may: a) use one word from the list of provided words, b) infer one word that is not in this list but which you believe better describes the topic; c) use a combination of a few words (i.e., a phrase) that describes the topic; d) explain the meaning of the topic in a sentence; e) some other method.

2. Please **organize** the topic labels you have chosen in part 1 above in the following manner:

1) place one topic label per line
2) rank these topic labels by relevance, where the first line is the most relevant to you, and the last line is the least relevant.

3. If you are **unable** to choose at least **one** topic label, please explain why.

The following Example 1 contains 4 topic labels, and the following example 2 provides an example answer of why no topic label could be identified:

Example 1

"word3
word1 word4
word k word m (inferred words)
Sentence explaining the meaning."

Example 2

"I was unable to choose a topic label because word1, word4, and word5 are about one topic, and word2, word6, and word10 were about a completely different topic. The rest of the words did not make any meaningful connection to me"

*

Your answer

Fig. 8 Survey v4 p2.

Fig. 9 Survey v4 p3.

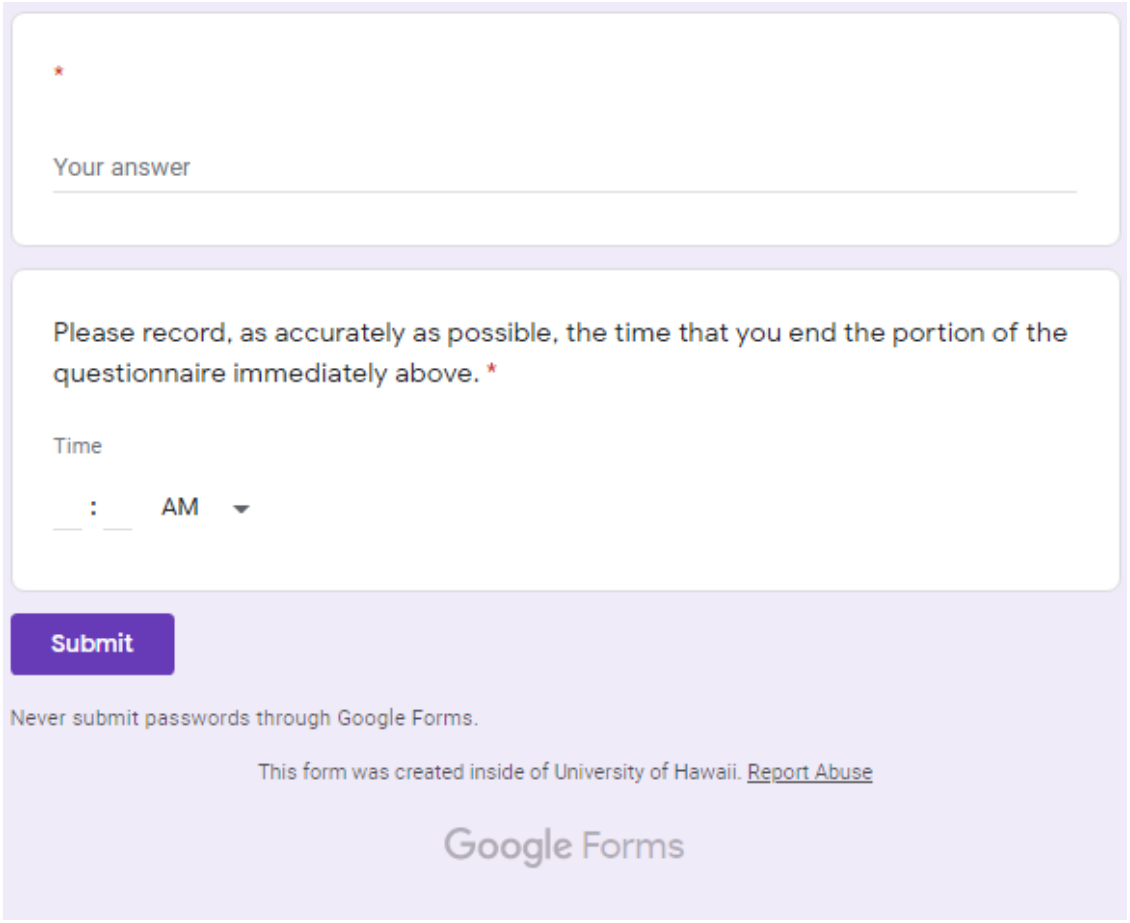


Fig. 10 Survey v4 p4.